DOCUMENT RESUME

ED 431 808 TM 029 927

AUTHOR Jiang, Hai
TITLE Estimation of Score Distributions for TOEFL Concordance Tables.
PUB DATE 1999-03-27
NOTE 19p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, Canada, April 20-22, 1999).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Computer Assisted Testing; *English (Second Language); Estimation (Mathematics); Scaling; *Scores; *Statistical Distributions; Tables (Data)
IDENTIFIERS *Concordance (Data); Paper and Pencil Tests; *Test of English as a Foreign Language

ABSTRACT

The purpose of this paper is to describe the techniques used in establishing the concordance tables between the Test of English as a Foreign Language (TOEFL), paper and pencil (P&P), and computer-based testing (CBT) sections and total reported score scales. Listening, reading, and composite structure and essay scores plus a total score are reported in the CBT on scales that are nonoverlapping with the P&P scales. The concordance study was conducted to make it possible to equate the CBT test to the P&P test. Examinees (n=8,387) took the CBT soon after their P&P tests. Responses from both were used to estimate the conditional distributions of the CBT section scores and the observed Essay scores, given the P&P scores. Projected population distributions were then obtained for the CBT scores and observed Essay scores. Given reference forms for the CBT, the CBT and Essay score distributions were projected to reference-test observed score distributions that then served as the basis for concordance functions. Concordance tables were successfully established between the CBT and the P&P reported score scales for TOEFL. It might be argued that the researchers could have used the study group alone to establish the concordance relationships. However, because of the self-selection of the examinees participating in the study, their P&P score distributions were clearly different from those of the population, and it was determined that more sophisticated methods needed to be employed. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 431 808

Estimation of Score Distributions for TOEFL Concordance Tables

Hai Jiang
Educational Testing Service

March 27, 1999

To be presented at the NCME Symposium, "Variations on Standard Computer Adaptive Testing Procedures", Montreal, Quebec, CANADA, April 20-22, 1999.

TM029927

BEST COPY AVAILABLE

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Hai Jiang
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Abstract

The purpose of this paper is to describe the techniques used in establishing the concordance tables between the TOEFL P&P and CBT section and total reported score scales. The TOEFL CBT consists of three sections Listening, Structure, and Reading just as in the P&P, plus a mandatory Essay session. Listening, Reading, and composite Structure and Essay scores plus a total score are reported for the CBT on scales that are non-overlapping with the P&P scales. Because of the new composite Structure and Essay score, it was unlikely that the Structure section and, hence, the total test could be equated to the P&P test, and hence a concordance study was conducted.

In the concordance study, examinees took the CBT soon after their P&P administrations. Using the responses on both the P&P and the CBT forms of the study group, we estimated the conditional distributions of the CBT section scores and the observed Essay scores, given the P&P scores. Using this distribution and the marginal distributions of the P&P section scores estimated from the national population, we then obtained projected population distributions for the CBT scores and observed Essay scores. Given reference forms for the CBT test, we projected the CBT and Essay score distributions to reference-test observed score distributions which then served as the basis for concordance functions.

Using the above techniques, we successfully established the concordance tables between the CBT and P&P reported score scales for TOEFL. One might argue that we could have used the study group alone to establish the concordance relationships. However, because of the self-selection of the examinees participated in the study, their P&P score distributions were clearly different from those of the population, and it was determined that more sophisticated methods needed to be employed.

Introduction

After several years of development, TOEFL introduced its CBT version to over 120 countries worldwide on July 24, 1998. The TOEFL CBT consists of three sections Listening (LC), Structure (ST), and Reading (RC) just as in the P&P, plus a mandatory Essay session. Both the Listening and the Structure sections are adaptive, but the Reading section is assembled linearly on the fly. Listening, Reading, and composite Structure and Essay scores plus a total score are reported for the CBT on scales that are non-overlapping with the P&P scales. Before the rollout of the TOEFL CBT, the Listening and the Structure reported scores ranged from 20 to 68, the Reading from 20 to 67, and the total from 200 to 677. The CBT scales are from 0 to 30 for the sections and from 0 to 300 for the total reported scores, and the P&P section and total reported score scales are now truncated at 31 and 310, respectively.

The CBT and the P&P tests of TOEFL are different not only in the ways they are given, but also in the measures themselves. There are many new item types for the CBT in the Listening and the Reading sections. For the Structure section, the reported score is now the composite of the Structure and the Essay scores.

Because of the differences in the measures, and in particular because of the new composite Structure and Essay score, it was unlikely that the sections and, hence, the total test could be equated to the P&P test, and hence a concordance study was conducted.

The Data

A total of 9,381 examinees from Nov 97, Dec 97, Jan 98, and Feb 98 TOEFL P&P administrations participated in the concordance study. They took the CBT soon after taking their P&P tests. After deleting records where the responses to any one of the sections were lacking or records that were duplicates, 9,247 of them remained. After matching these records with their P&P records, a total of 860 examinees were further removed because of the reasons listed below:

- Their records were not used in P&P calibration;
- Their CBT and P&P records could not be matched;
- Their P&P records contained no valid reported scores.

Motivation study

For the remaining 8,387 examinees, their CBT scores (section estimated θs) were compared with their P&P scores (section reported scores) to flag those who did exceptionally poor or well on the CBT. However, remember that the CBT section θs and the P&P reported scores are not directly comparable. A remedy of this problem is to use the percentile ranks of the scores (the rank of the CBT θ and that of the P&P reported score) because they are on the same scale and have the same Uniform distribution. Thus, the flagging of examinees who did exceptionally poor (dubbed under-motivated) or well (over-motivated) on the CBT was based on the following criteria:

$$|R_{CBT} - R_{P\&P}| \geq c_R \text{ or}$$

$$|L_{CBT} - L_{P\&P}| \geq c_L$$

where R_{CBT} and $R_{P\&P}$ are the percentile ranks of the CBT θ and the P&P reported score of an examinee: $L = \ln\frac{R}{1-R}$ is the logit of the percentile rank R and has a Normal distribution. While the constant c_L here was set at twice the S.D. of $L_{CBT} - L_{P\&P}$, c_R was set at 0.15 (more or less arbitrarily) to exclude people whose percentile rankings on the

CBT and the P&P were too different (especially at the two extreme ends of the ranking scale).

Figure 1 below shows the plot of the logits of CBT score rankings vs. those of the P&P score rankings for the LC measure. In Figure 1, the x-axis labeled TOEFL is the logit of P&P reported score ranking, and the y-axis labeled CONCORD is the logit of CBT θ ranking. The points outside the region bounded by the two lines are those flagged as having inconsistent performances on the CBT and the P&P LC measures. The patterns here are mainly caused by the discrete nature of the P&P reported score distribution.

Insert Figure 1 about here

Records of the 834 examinees who were flagged for any of the sections were deleted, leaving 7,553 records with valid and consistent scores on all the sections of the CBT as well as the P&P tests. These records were then matched with their CBT Essay records, and 496 of them could not be matched leaving us 7,057 complete records with P&P section responses, CBT section responses, and CBT Essay scores.

A sample (the national sample) of 50,000 examinees representing the national population was assembled using the examinees from the same four P&P administrations. This sample had the above 7,057 examinees as a subset. Each examinee in the national sample had only P&P section responses.

The Analysis

Step 1

Remember that we have item parameter estimates for the four P&P administrations and parameter estimates for the CBT items (they are all on the same P&P scale). Using the MGROUP estimation programs and assuming that the latent variables are Multivariate Normal, we got the estimated $\boldsymbol{\theta}_{P\&P}$ distribution for the national sample using their P&P section responses. MGROUP is a set of computer programs, each of which gives maximum likelihood estimates of $\boldsymbol{\gamma}$ and $\boldsymbol{\Sigma}$ in the regression model:

$$\boldsymbol{\theta} = \mathbf{X}^T\boldsymbol{\gamma} + \mathbf{e}$$

where $\mathbf{e}$ is Multivariate Normal with Mean $\mathbf{0}$ and Covariance $\boldsymbol{\Sigma}$, $\boldsymbol{\theta}$ is a multidimensional latent variable and $\mathbf{X}$ is a user-specified design matrix. Each examinee's vector of conditioning variables can serve as a row in the design matrix. Numerical approximation of a solution based on the E-M algorithm is used (Mislevy, 1985). Imputed ability estimates can be output, which are not "examinee optimal", but designed to provide consistent estimates of subgroup distributions.

Again using MGROUP and this time conditioning on the CBT Essay scores, we got the estimated $(\boldsymbol{\theta}_{P\&P}, \boldsymbol{\theta}_{CBT} \mid \text{Essay})$ distribution for the sample of 7,057 examinees with P&P, CBT, and Essay records. Among these 7,057 examinees, 501 of them with Essay score 0 were removed because we could not tell if their Essays were off-topic or blank due to a software problem[1]. For the 6,556 examinees left (the concordance sample), we had the $(\boldsymbol{\theta}_{P\&P}, \boldsymbol{\theta}_{CBT} \mid \text{Essay})$ distribution. Along with the Essay score distribution, we got the joint distribution of $(\boldsymbol{\theta}_{P\&P}, \boldsymbol{\theta}_{CBT}, \text{Essay})$ for the concordance sample.

[1] For some examinees who chose to key enter their Essays, their files might not be saved when their allotted time expired. This problem has since been corrected.

Step 2

For the concordance sample, if the distribution of $(\boldsymbol{\theta}_{P\&P}, \boldsymbol{\theta}_{CBT})$ is Multivariate Normal with Mean $\boldsymbol{\mu} = (\boldsymbol{\mu}_{P\&P}, \boldsymbol{\mu}_{CBT})^T$ and Covariance $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{P\&P} & \boldsymbol{\Sigma}_{CBT.P\&P}^T \\ \boldsymbol{\Sigma}_{CBT.P\&P} & \boldsymbol{\Sigma}_{CBT} \end{pmatrix}$, the distribution of $(\boldsymbol{\theta}_{CBT} \mid \boldsymbol{\theta}_{P\&P} = \mathbf{x}_{P\&P})$ is also Multivariate Normal with Mean $\boldsymbol{\mu}_{CBT} + \boldsymbol{\Sigma}_{CBT.P\&P}\boldsymbol{\Sigma}_{P\&P}^{-1}(\mathbf{x}_{P\&P} - \boldsymbol{\mu}_{P\&P})$ and Covariance $\boldsymbol{\Sigma}_{CBT} - \boldsymbol{\Sigma}_{CBT.P\&P}\boldsymbol{\Sigma}_{P\&P}^{-1}\boldsymbol{\Sigma}_{CBT.P\&P}^T$.

To show that the distribution of $(\boldsymbol{\theta}_{P\&P}, \boldsymbol{\theta}_{CBT})$ is indeed Multivariate Normal, we can show that the distributions of $\delta_{P\&P\,LC}$, $(\delta_{P\&P\,SAWE} \mid \delta_{P\&P\,LC})$, $(\delta_{P\&P\,RC} \mid \delta_{P\&P\,LC}, \delta_{P\&P\,SAWE})$, $(\delta_{CBT\,LC} \mid \boldsymbol{\theta}_{P\&P})$, $(\delta_{CBT\,ST} \mid \boldsymbol{\theta}_{P\&P}, \delta_{CBT\,LC})$, and $(\delta_{CBT\,RC} \mid \boldsymbol{\theta}_{P\&P}, \delta_{CBT\,LC}, \delta_{CBT\,ST})$ are successively Normal.

To show that the distribution of $(\delta_{CBT\,ST} \mid \boldsymbol{\theta}_{P\&P}, \delta_{CBT\,LC})$ is Normal, for example, we only need to show that the residuals of the linear regression of $\delta_{CBT\,ST}$ on $\boldsymbol{\theta}_{P\&P}$ and $\delta_{CBT\,LC}$ have a Normal distribution. This can be shown by the QQ Plot of the residuals (Plot of the quantiles of the residuals against those of the Standard Normal distribution).

Figure 2 below gives the QQ Plot of the residuals of the linear regression of $\delta_{CBT\,ST}$ on $\boldsymbol{\theta}_{P\&P}$ and $\delta_{CBT\,LC}$. The plot is very close to being a straight line, indicating that the distribution of the residuals resembles closely that of a Normal distribution.

Insert Figure 2 about here

Now that we have shown the distribution of $(\boldsymbol{\theta}_{P\&P}, \boldsymbol{\theta}_{CBT})$ for the concordance sample is Multivariate Normal, its Mean and Covariance can be estimated by the sample Mean $\mathbf{m} = (\mathbf{m}_{P\&P}, \mathbf{m}_{CBT})^T = ((0.4574, 0.4063, 0.4343), (0.4763, 0.3406, 0.5577))^T$ and the sample Covariance

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{P\&P} & \mathbf{S}_{CBT.P\&P}^T \\ \mathbf{S}_{CBT.P\&P} & \mathbf{S}_{CBT} \end{pmatrix} = \left(\begin{pmatrix} 0.5887 & 0.4697 & 0.4535 \\ 0.4697 & 0.7668 & 0.6302 \\ 0.4535 & 0.6302 & 0.6379 \end{pmatrix} \begin{pmatrix} 0.5977 & 0.4455 & 0.4055 \\ 0.5008 & 0.7109 & 0.5780 \\ 0.5020 & 0.5894 & 0.5837 \end{pmatrix} \right.$$
$$\left. \begin{pmatrix} 0.5977 & 0.5008 & 0.5020 \\ 0.4455 & 0.7109 & 0.5894 \\ 0.4055 & 0.5780 & 0.5837 \end{pmatrix} \begin{pmatrix} 0.6607 & 0.4867 & 0.4752 \\ 0.4867 & 0.6677 & 0.5559 \\ 0.4752 & 0.5559 & 0.5787 \end{pmatrix} \right)$$

Since for the concordance sample $(\boldsymbol{\theta}_{P\&P}, \boldsymbol{\theta}_{CBT})$ has a Multivariate Normal distribution, for any given $\mathbf{x}_{P\&P}$, $(\boldsymbol{\theta}_{CBT} \mid \boldsymbol{\theta}_{P\&P} = \mathbf{x}_{P\&P})$ is also Multivariate Normal with Mean $\begin{pmatrix} -0.0035 + 0.9126x_{P\&P\,LC} - 0.1035x_{P\&P\,ST} + 0.2405x_{P\&P\,RC} \\ -0.0445 + 0.0271x_{P\&P\,LC} + 0.8875x_{P\&P\,ST} + 0.0280x_{P\&P\,RC} \\ 0.1662 - 0.0368x_{P\&P\,LC} + 0.0149x_{P\&P\,ST} + 0.9264x_{P\&P\,RC} \end{pmatrix}$ and Covariance $\begin{pmatrix} 0.0464 & 0.0121 & 0.0246 \\ 0.0121 & 0.0083 & 0.0156 \\ 0.0246 & 0.0156 & 0.0442 \end{pmatrix}$.

The distribution of $(\text{Essay} \mid \boldsymbol{\theta}_{P\&P}, \boldsymbol{\theta}_{CBT})$ for the concordance sample can be estimated using logistic regression. Specifically, the logit of $P(\text{Essay} > e)$ is linearly regressed on $\boldsymbol{\theta}_{P\&P}$ and $\boldsymbol{\theta}_{CBT}$ for e=1,2,....,5, consecutively using the concordance sample data. The results are given below:

$$\begin{aligned}\text{logit}(P(\text{Essay} > 1)) = {} & 2.6455 - 4.3540\theta_{P\&P\,LC} + 46.7614\theta_{P\&P\,ST} - 17.7988\theta_{P\&P\,RC} \\ & + 7.3826\theta_{CBT\,LC} - 49.2849\theta_{CBT\,ST} + 17.3622\theta_{CBT\,RC}\end{aligned}$$

$$\begin{aligned}\text{logit}(P(\text{Essay} > 2)) = {} & 2.7682 - 0.3298\theta_{P\&P\,LC} - 3.2557\theta_{P\&P\,ST} - 0.4054\theta_{P\&P\,RC} \\ & + 0.8197\theta_{CBT\,LC} + 4.5897\theta_{CBT\,ST} + 0.1396\theta_{CBT\,RC}\end{aligned}$$

$$\begin{aligned}\text{logit}(P(\text{Essay} > 3)) = {} & 0.5790 + 1.1004\theta_{P\&P\,LC} - 4.6485\theta_{P\&P\,ST} + 0.9660\theta_{P\&P\,RC} \\ & - 0.6526\theta_{CBT\,LC} + 6.1362\theta_{CBT\,ST} - 1.2172\theta_{CBT\,RC}\end{aligned}$$

$$\begin{aligned}\text{logit}(P(\text{Essay} > 4)) = {} & -3.4033 - 1.6721\theta_{P\&P\,LC} + 11.1888\theta_{P\&P\,ST} - 3.1080\theta_{P\&P\,RC} \\ & + 3.0829\theta_{CBT\,LC} - 10.7002\theta_{CBT\,ST} + 2.2180\theta_{CBT\,RC}\end{aligned}$$

$$\begin{aligned}\text{logit}(P(\text{Essay} > 5)) = {} & -8.4057 - 4.3671\theta_{P\&P\,LC} + 26.5182\theta_{P\&P\,ST} - 6.2744\theta_{P\&P\,RC} \\ & + 7.0661\theta_{CBT\,LC} - 26.6417\theta_{CBT\,ST} + 4.3429\theta_{CBT\,RC}\end{aligned}$$

After having estimated the distributions of $(\boldsymbol{\theta}_{CBT} \mid \boldsymbol{\theta}_{P\&P})$ and $(\text{Essay} \mid \boldsymbol{\theta}_{P\&P}, \boldsymbol{\theta}_{CBT})$ for the concordance sample, we got the distribution of $(\boldsymbol{\theta}_{CBT}, \text{Essay} \mid \boldsymbol{\theta}_{P\&P})$. Assuming this distribution is the same for both the concordance and the national samples, we can get the joint distribution of $(\boldsymbol{\theta}_{P\&P}, \boldsymbol{\theta}_{CBT}, \text{Essay})$ for the national sample by using the distributions of $(\boldsymbol{\theta}_{CBT}, \text{Essay} \mid \boldsymbol{\theta}_{P\&P})$ and $\boldsymbol{\theta}_{P\&P}$ for the national sample. For example, for a given point $\mathbf{x}_{P\&P} = (0.4340, 0.3599, 0.0393)^T$ from the $\boldsymbol{\theta}_{P\&P}$ distribution for the national sample, we drew a value $\mathbf{x}_{CBT} = (0.5296, 0.2829, 0.2502)^T$ from the distribution of $(\boldsymbol{\theta}_{CBT} \mid \boldsymbol{\theta}_{P\&P} = \mathbf{x}_{P\&P})$ which is Multivariate Normal with Mean $(0.3647, 0.2878, 0.1920)^T$ and Covariance $\begin{pmatrix} 0.0464 & 0.0121 & 0.0246 \\ 0.0121 & 0.0083 & 0.0156 \\ 0.0246 & 0.0156 & 0.0442 \end{pmatrix}$. For $(\mathbf{x}_{P\&P}, \mathbf{x}_{CBT})$, a value 3 for the Essay score was then drawn using the distribution of $(\text{Essay} \mid \boldsymbol{\theta}_{P\&P}, \boldsymbol{\theta}_{CBT})$ which is given by $P(\text{Essay} \le e) = 0.00001, 0.03898, 0.37582, 0.74340, 0.97615, 1.00000$ for $e = 1, 2, \ldots, 6$.

Step 3

Reference forms for the CBT sections were assembled linearly. These forms used the same types of items and conformed to the same sets of content specifications used in actual CBT tests (adjusted for test length if necessary). The CBT ST section uses the same reference form as used in the P&P test that consists of 38 items because the same item types are used in both the P&P and the CBT tests. The reference forms for the CBT LC and the RC sections consist of 50 and 44 items (compared to 50 and 49 items for the P&P LC and the RC sections), respectively and have similar psychometric characteristics of the ones for the corresponding P&P sections.

For given values in the $(\boldsymbol{\theta}_{P\&P}, \boldsymbol{\theta}_{CBT})$ space, using the P&P and the CBT reference form item parameters we can simulate the observed item response vectors to the P&P and the CBT tests (and thus the observed P&P and CBT section scores). Using resampling (sampling with replacement) from the joint distribution of $(\boldsymbol{\theta}_{P\&P}, \boldsymbol{\theta}_{CBT}, \text{Essay})$ for the national sample, we can then estimate their observed P&P and CBT section score and Essay score distributions.

Step 4

Recall that for the ST section in the CBT a composite score incorporating examinee performances on the Structure and on Essay is reported. In deriving the CBT ST composite score, we decided to give equal weight to Essay and the ST observed scores (adjusted for scale differences in Essay and the ST observed scores) using the following formula:

$$C_{CBT\,ST} = X_{CBT\,ST} + \frac{\hat{\sigma}(X_{CBT\,ST})}{\hat{\sigma}(\text{Essay})} \cdot \text{Essay}$$

here $X_{CBT\,ST}$ is the ST observed score, $\hat{\sigma}(X_{CBT\,ST})$ and $\hat{\sigma}(\text{Essay})$ are the S.D. estimates of the CBT ST observed and Essay scores, respectively. From the distributions of the CBT ST observed scores and Essay scores, we have $\hat{\sigma}(X_{CBT\,ST}) = 6.4990$ and $\hat{\sigma}(\text{Essay}) = 0.9495$.

Once we have determined how to compute the CBT ST composite score, its distribution can be estimated again using resampling techniques as in Step 3. This distribution along with the distributions of the P&P section observed and the CBT LC and the RC section observed scores then served as the basis for the concordance functions.

The CBT total reported score is computed as ten-thirds of the sum of the LC, ST composite and RC reported scores just as in the P&P test.

To convert the CBT observed section scores to the reported score scale, we need to first define the section reported score distribution as follows:

- Standardize the observed LC, ST composite, and RC scores to the range from 0 to 30;
- Get total scores as ten-thirds of the sum of these standardized scores;
- Obtain the distribution of the total scores and smooth it using a Negative Hypergeometric distribution;
- Scale back this distribution to the range from 0 to 30.

Figure 3 below shows the reported score distribution defined in the above manner. As is clear from Figure 3, this distribution is skewed to the left, which is commonly seen in distributions of test scores. It combines the characteristics of CBT observed section score distributions in a direct manner, and resembles the shape of the P&P total reported score distribution. The section reported scores with the above distribution have a Mean of 19.7 and S.D. of 4.9, and the total reported scores have a Mean of 195.1 and S.D. of 44.5.

Insert Figure 3 about here

Step 5

The conversion curves between CBT section observed and reported scores can be obtained by equipercentile equating using the CBT section reported score distribution obtained in Step 4, and the distributions of the CBT LC observed, ST composite, and RC observed scores.

The P&P section observed scores are converted into reported scores using the conversion curves for the reference forms.

Once we have obtained the section reported scores for the P&P and the CBT tests, we can get the frequencies of the P&P and the CBT total reported scores. Along with the frequencies of the P&P and the CBT section observed scores, we get a set of four concordance tables using equipercentile equating:

P&P LC observed to CBT LC observed,
P&P S&WE observed to CBT ST composite,
P&P RC observed to CBT RC observed, and
P&P total reported to CBT total reported.

Figure 4 gives the plots of the concordance relationships between the P&P and the CBT tests for section observed and total reported scores. Except for the curve between the P&P ST observed scores and the CBT ST composite scores, these relationships have only minor curvatures throughout their range. The sudden change of curvature in the upper end of the concordance relationship between the P&P ST observed and the CBT ST composite scores is caused by the differences in the distributions of the two scores. Figure 5 gives the distributions of these two scores. The shape of the P&P ST observed score distribution is commonly seen. However, the distribution of the CBT ST composite scores exhibits unusual patterns caused by the combination of two totally different distributions. The Essay score distribution is more or less symmetric with most of the masses concentrating in the middle (at 3 or 4 point) while the CBT ST observed score distribution is similar in shape to that of the P&P ST observed scores. The combination of these two different distributions thus produces multi-modes in the distribution of the CBT ST composite scores as seen in Figure 5.

Insert Figures 4 and 5 about here

Using the conversion curves for the P&P sections, the concordance tables between the P&P and the CBT section observed scores, and the conversion curves for the CBT sections, we can get concordance tables between the P&P and the CBT section reported scores. Specifically, for a given P&P section reported score $S_{P\&P}$, using the conversion curve for the P&P section and interpolation, we get the corresponding P&P observed section score $X_{P\&P}$. Using the concordance table and interpolation, we get an equivalent CBT observed section score X_{CBT}, which is then converted into the reported score S_{CBT} :

$$S_{P\&P} \rightarrow X_{P\&P} \leftrightarrow X_{CBT} \rightarrow S_{CBT}$$

Figure 6 below shows the concordance curves between the P&P and the CBT section reported scores. The patterns seen in the concordance between the P&P ST observed and the CBT ST composite scores carry over here. The concordance curve for the LC reported scores has change of steepness in the lower end, which is caused by the linear portion (below chance level) of the conversion curve for the P&P observed scores.

Insert Figure 6 about here

Summary

Using the techniques described above, we successfully established the concordance tables between the CBT and P&P reported score scales for TOEFL which are shown in Tables 1 to 4 below.

Insert Tables 1 to 4 about here

One might argue that we could have used the study group alone to establish the concordance relationships. However, because of the self-selection of the examinees participated in the study, their P&P score distributions were clearly different from those of the population, and it was determined that something more sophisticated needed to be employed. Table 5 gives the Means of the P&P section observed scores and total reported scores for the concordance sample as well as for the national sample.

Insert Table 5 about here

As can be seen from table 5, because of self-selection, examinees who actually participated tended to be more able than those in the population. Thus complex mechanisms like those in the previous section are needed to reduce this artificial effect.

Acknowledgement

The work described in this paper was done by the collaborative efforts of many people at Educational Testing Service. In particular, I want to thank Robert Mislevy and Daniel Eignor for their theoretical guidance, and Norma Norris for her support for running the MGROUP programs. I also want to thank Walter "Denny" Way, Samuel Livingston, Neil Dorans, and Patricia Carey among others for their numerous comments, and suggestions, as well as criticisms.

Reference

Mislevy, R.J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80*, 993-997.

Figure 1. Plot of the logits of CBT score rankings
vs. those of the P&P score rankings
The LC measure.

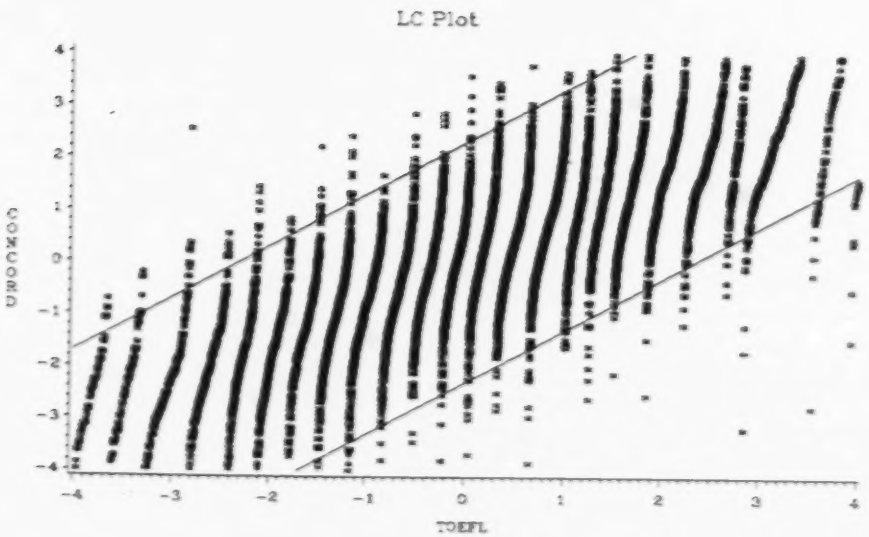


Figure 2. QQ Plot of residuals of CBT ST on P&P and CBT LC

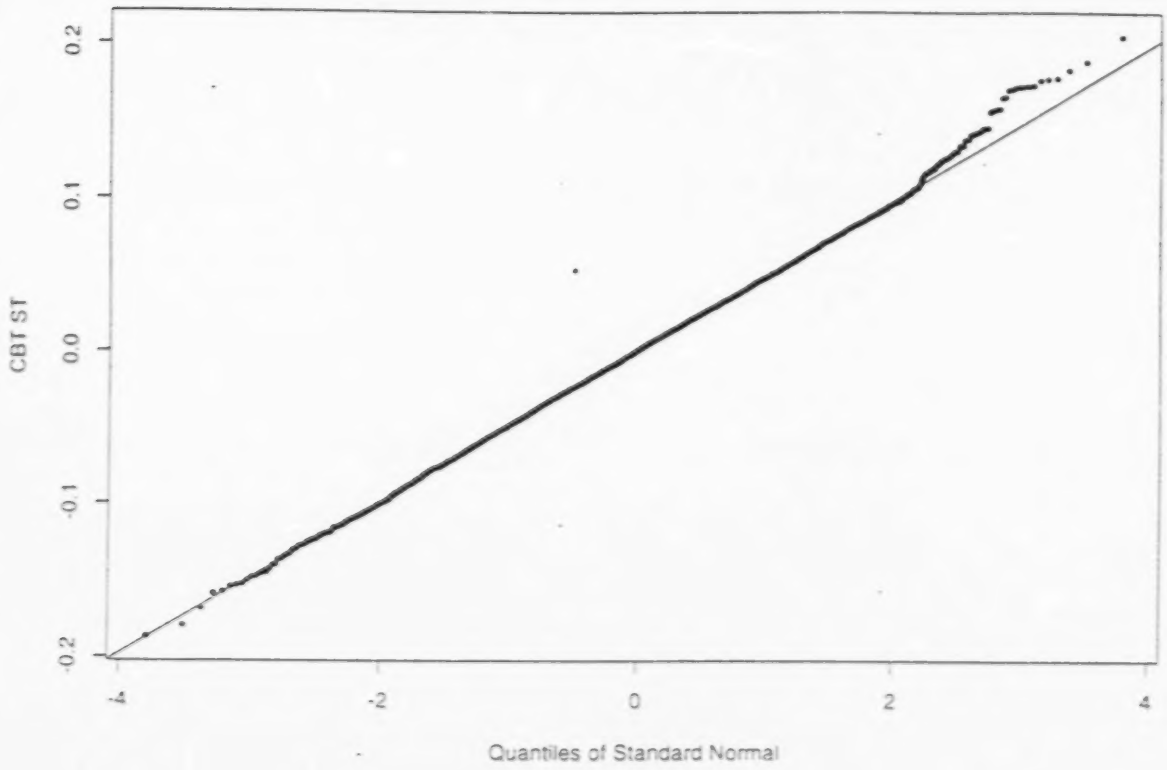


Figure 3. CBT section reported score distribution.

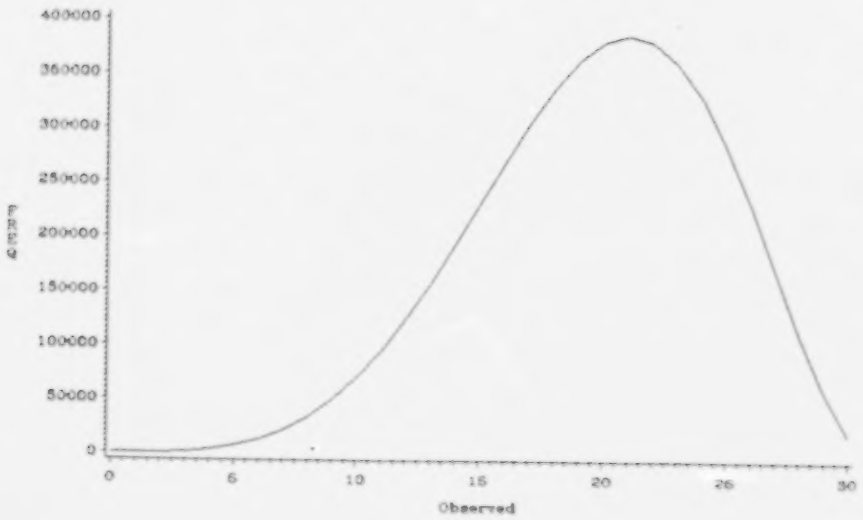
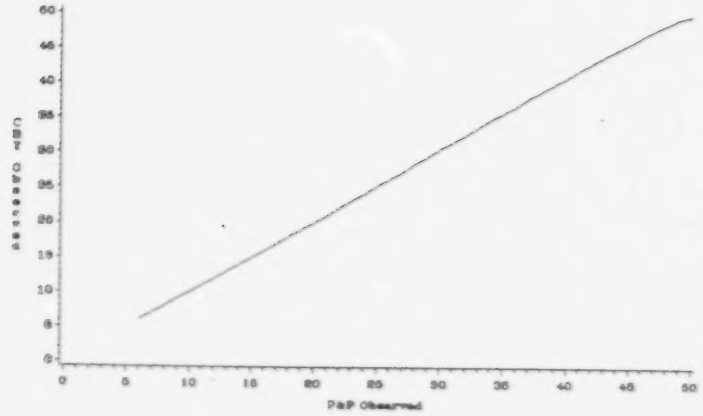
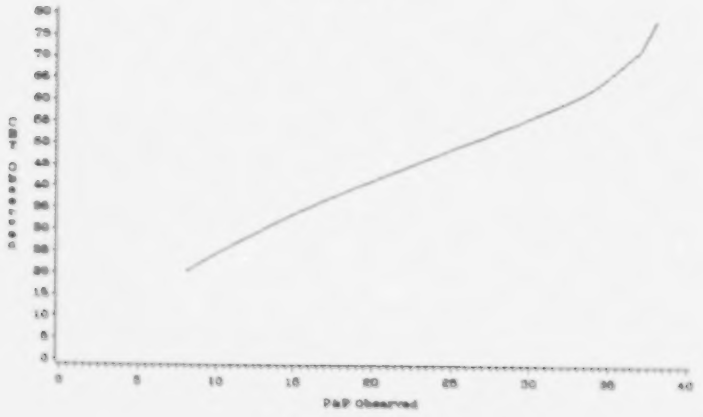


Figure 4. Plots of concordance curves for section observed and total reported scores

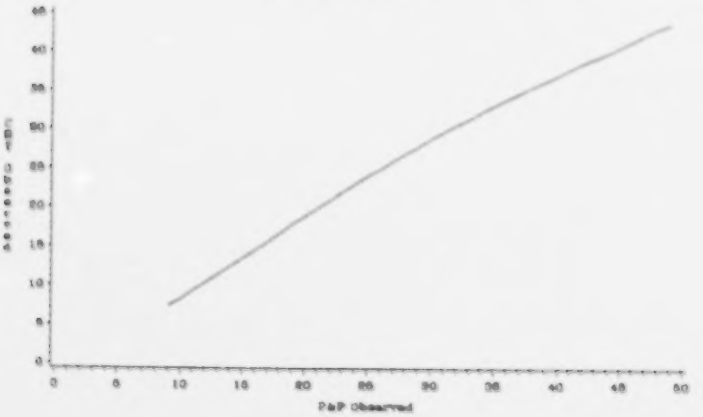
LC section



ST section

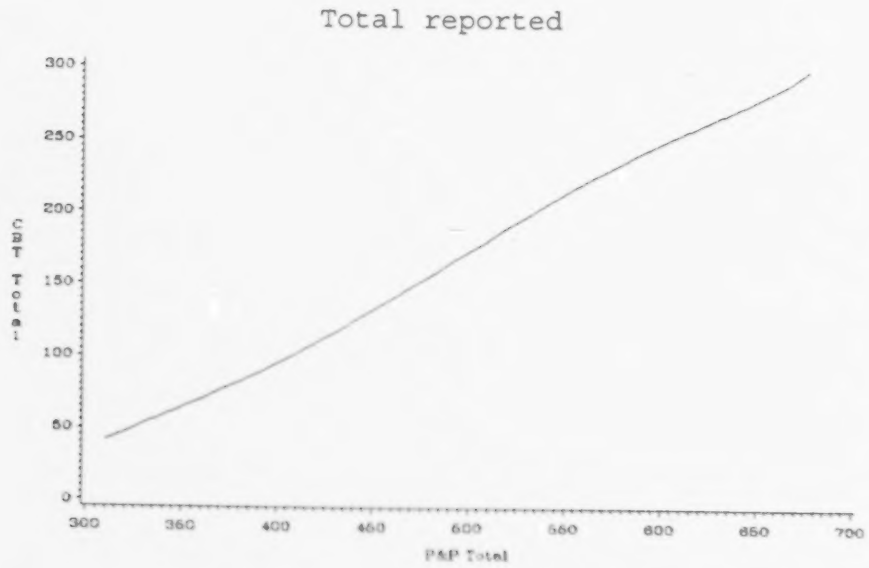


RC section



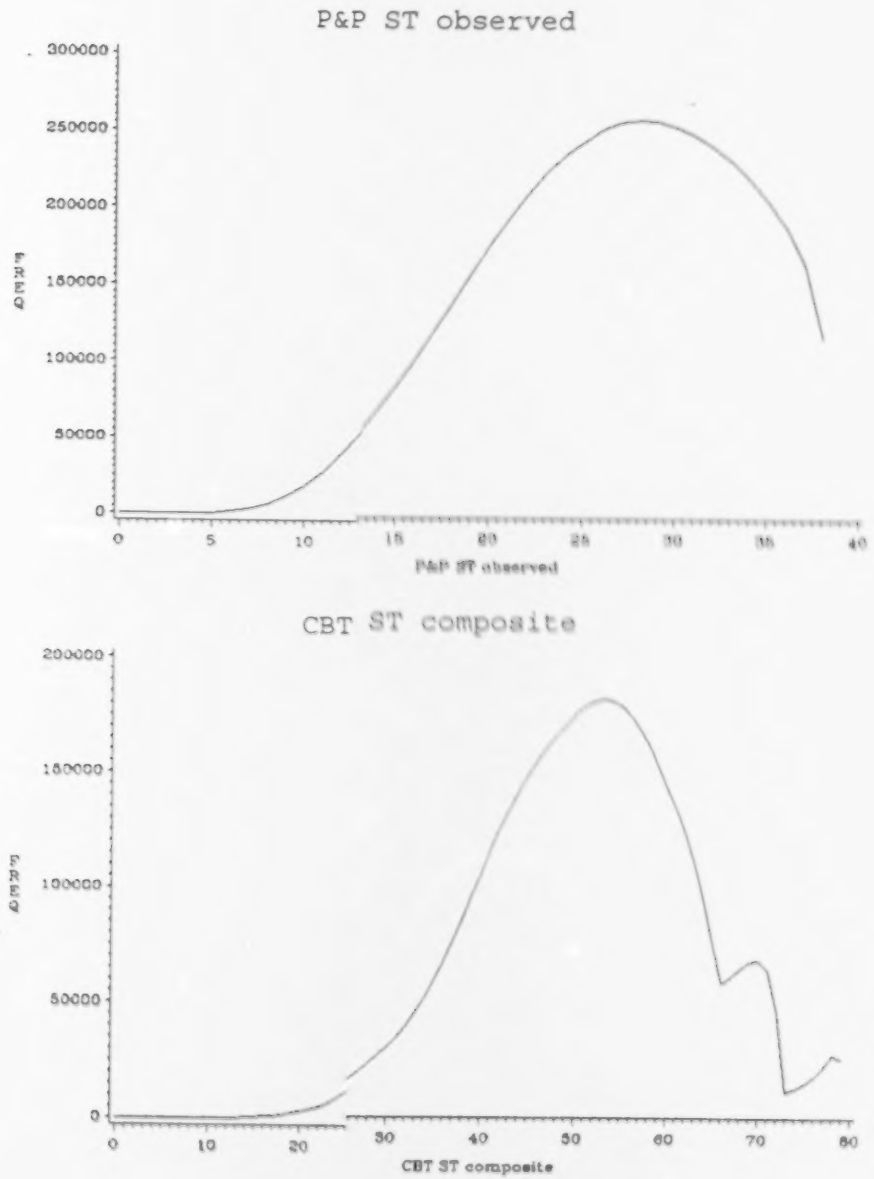
BEST COPY AVAILABLE

Figure 4. Plots of concordance curves for section observed and total reported scores (cont.)



BEST COPY AVAILABLE

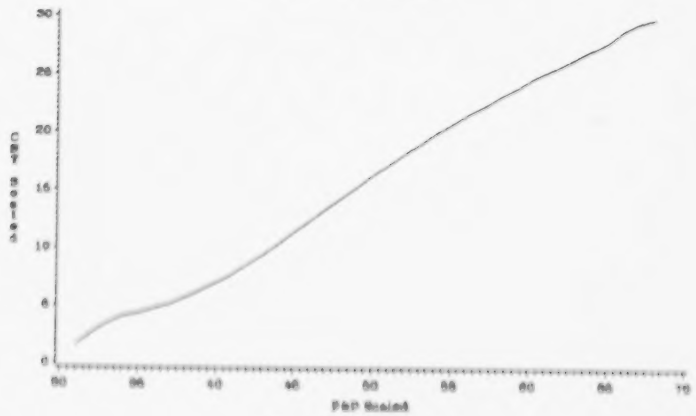
Figure 5. Distributions of the P&P ST observed scores and the CBT ST composite scores



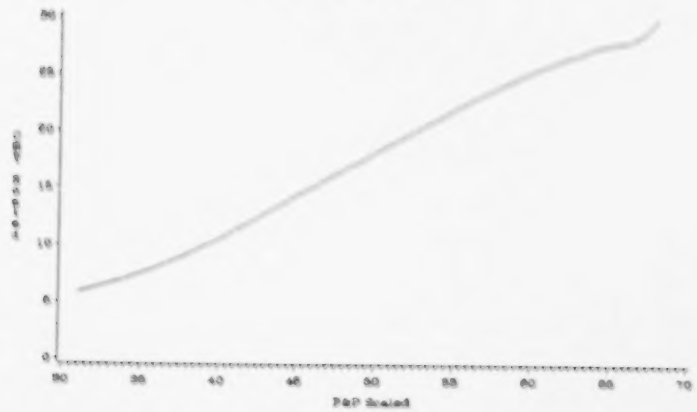
BEST COPY AVAILABLE

Figure 6. Plots of concordance curves for section reported scores

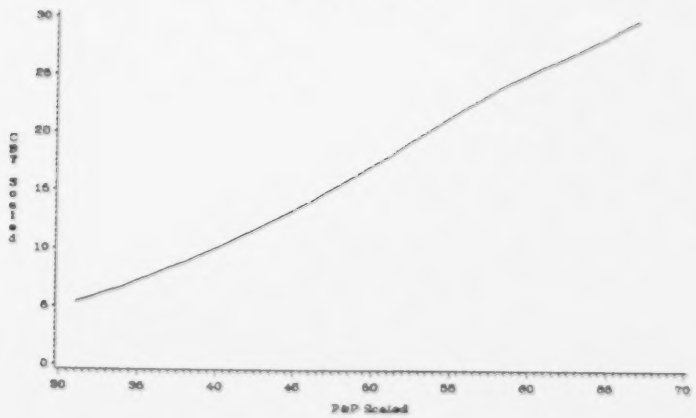
LC section



ST section



RC section



BEST COPY AVAILABLE

Table 1. Concordance between section reported scores

Listening

P&P	CBT	P&P	CBT	P&P	CBT	P&P	CBT
68	30	58	23	48	14	38	6
67	30	57	22	47	13	37	5
66	29	56	22	46	12	36	5
65	28	55	21	45	11	35	4
64	27	54	20	44	10	34	4
63	27	53	19	43	9	33	3
62	26	52	18	42	9	32	3
61	25	51	17	41	8	31	2
60	25	50	16	40	7		
59	24	49	15	39	6		

Table 2. Concordance between section reported scores

Structure

P&P	CBT	P&P	CBT	P&P	CBT	P&P	CBT
68	30	58	24	48	17	38	9
67	29	57	23	47	16	37	9
66	28	56	23	46	15	36	8
65	28	55	22	45	14	35	8
64	27	54	21	44	14	34	7
63	27	53	20	43	13	33	7
62	26	52	20	42	12	32	6
61	26	51	19	41	11	31	6
60	25	50	18	40	11		
59	25	49	17	39	10		

Table 3. Concordance between section reported scores

Reading

P&P	CBT	P&P	CBT	P&P	CBT	P&P	CBT
		58	24	48	16	38	9
67	30	57	23	47	15	37	8
66	29	56	22	46	14	36	8
65	28	55	21	45	13	35	7
64	28	54	21	44	13	34	7
63	27	53	20	43	12	33	6
62	26	52	19	42	11	32	6
61	26	51	18	41	11	31	5
60	25	50	17	40	10		
59	25	49	16	39	9		

Table 4. Concordance between total reported scores

P&P	CBT	P&P	CBT	P&P	CBT	P&P	CBT
677	300	583	237	490	163	397	93
673	297	580	237	487	163	393	90
670	293	577	233	483	160	390	90
667	290	573	230	480	157	387	87
663	287	570	230	477	153	383	83
660	287	567	227	473	150	380	83
657	283	563	223	470	150	377	80
653	280	560	220	467	147	373	77
650	280	557	220	463	143	370	77
647	277	553	217	460	140	367	73
643	273	550	213	457	137	363	73
640	273	547	210	453	133	360	70
637	270	543	207	450	133	357	70
633	267	540	207	447	130	353	67
630	267	537	203	443	127	350	63
627	263	533	200	440	123	347	63
623	263	530	197	437	123	343	60
620	260	527	197	433	120	340	60
617	260	523	193	430	117	337	57
613	257	520	190	427	113	333	57
610	253	517	187	423	113	330	53
607	253	513	183	420	110	327	50
603	250	510	180	417	107	323	50
600	250	507	180	413	103	320	47
597	247	503	177	410	103	317	47
593	243	500	173	407	100	313	43
590	243	497	170	403	97	310	40
587	240	493	167	400	97		

Table 5. Means of the P&P section observed scores and total reported scores

Sample	P&P LC	P&P ST	P&P RC	Total
National	35.6	26.6	33.5	528.7
Concord	36.6	28.0	35.9	541.3